# Automatic mapping of urban wastewater networks based on manhole cover locations

Nanée Chahinian, Carole Delenne, Benjamin Commandre, Mustapha Derras, Laurent Deruelle, Jean-Stéphane Bailly

# Automatic mapping of urban wastewater networks based on manhole cover locations

Nanée Chahinian[a], Carole Delenne[a,b], Benjamin Commandré[a,b], Mustapha Derras[c], Laurent Deruelle[c], Jean-Stéphane Bailly[d,e]

[a] *HSM, Univ. Montpellier, CNRS, IRD, Montpellier, France*
[b] *Inria Lemon, CRISAM - Inria Sophia Antipolis - Méditerranée, France*
[c] *Berger-Levrault, Montpellier, France*
[d] *LISAH, Univ Montpellier, INRA, IRD, Montpellier SupAgro, Montpellier, France*
[e] *AgroParisTech, Paris, France*

**Abstract**

Accurate maps of wastewater networks in cities are mandatory for an integrated management of water resources. However, in many countries around the world this information is unavailable or inaccurate. A new mapping method is put forward to create wastewater network maps using manhole cover locations as a prime information source. These locations could be available via ground surveys, remote sensing techniques or stakeholder's databases. A new algorithm is developed which considers manhole covers as the nodes of the network and connects them automatically. It minimizes cost functions defined by industry rules thus generating an optimized network. The various input data and the rules used to build the deterministic tree-shaped graph being uncertain, a stochastic version of the algorithm is also put forward to generate a set of probable networks in addition to the optimized one. The method is tested on the wastewater networks of Prades-le-Lez and validated on the town of Ramonville Saint Agne. Both towns are located in Southern France, are part of the two most dynamic metropolitan areas of France and are under constant urban pressure due to their proximity with the cities of Montpellier and Toulouse. The shape and topology of the mapped networks are compared to the actual ones. The results indicate an overall good agreement between the layouts of the real and generated

networks. The proposed algorithm may thus be used to map wastewater networks from sampled georeferenced manhole covers, elevation and street network databases. Additional sources of information are however necessary to recreate the networks full geometry and insure proper conveyance. The low error values and high scores for completeness, correctness and quality indicate that the method is robust and may be adapted and tested on other study zones.

*Keywords:* Wastewater, Urban hydrology, Multigraph, Network connectivity, Mapping

---

## 1. Introduction

Urban expansion is an ongoing process both in developing and developed countries. According to the latest global statistics published by the UN 55% of the world's population is currently residing in urban areas as opposed to 30% in 1950. By 2050 this percentage is estimated to reach 68% [35]. The increase in population often leads to urban sprawl and city managers have to constantly extend water access and sanitation services to new peripheral areas. Getting accurate and updated information on the underground wastewater and stormwater networks is a cumbersome task, especially in cities undergoing urban expansion [24, 30]. With the development of smart city technologies, a growing number of towns are getting equipped with electronic sensors which are able to log and transmit data continuously. It thus becomes easier to monitor the evolution of environmental variables in quasi real time and to make adequate decisions based on the prevailing conditions. However, data on the location and the geometric features of water networks are still incomplete or missing [10]. This is especially due to the fact that many countries only recently passed bills on the localization accuracy and precision that contractors need to report back when they undertake works in the vicinity of underground utility networks. In France for instance, this bill was passed in 2012 [22].

Attempts have been made to find the optimal design configuration for wastewater and stormwater networks based on topographic and hydraulic constraints or economic costs [40, 2, 21, 27, 31]. Network layout optimisation problems have been thoroughly investigated in operational research since the 1950s [16, 7]. In combinartorial optimisation this equates to finding the minimum spanning tree of an undirected, connected graph and is known as such as put forward by Borvka in 1926 [32] and later by Kruskal and Prim [23, 36] (see a review in [17, 7]) or finding the shortest path to connect N-points commonly known as the Steiner tree problem [19, 13].

Some authors have generated virtual urban networks [26, 39]. However, in applications where the actual network has to be mapped, exogenous factors such as existing network branches may render these solutions quite inefficient. Fewer attempts have been made in the literature to reproduce the layout of actual drainage networks, on urban or farmed catchments [6]. In [4], [8] and [11] the node to link information was provided by urban databases or local operators. In the case of [10], who have developed a Bayesian Mapping model for buried utility networks using utility records coupled with MTU sensor data and manhole surveys, the node to link connection was established through hypothesis testing using the Expectation Maximization Algorithm. Their results were validated against 2 site specific data.

When working on an urban catchment where very little information about the network configuration is available, one solution may be to infer it based on visible features such as manhole covers. Recent works have shown that these could be localized by using new processing techniques and high resolution aerial imagery [29, 33, 34]. Although the precise location of all the manhole covers is not mandatory to build accurate hydraulic models of the wastewater network, this is far from being the only piece of information required by modelling soft-

ware products. The conveyance should be assessed via parameters such as pipe size, shape and roughness. Pipe slope is also a parameter of great importance as classical hydraulic models cannot compute gravity fed flow on counter-slopes, unless pumping stations are installed.

In this paper, we put forward a methodology to generate a map of a wastewater or a stormwater network based on manhole cover or sewer inlet locations. The paper is structured as follows: the optimization algorithm and the cost functions are presented in section 2.1. Section 2.2 presents the stochastic version of this algorithm, allowing input data and rules uncertainty propagation up to the generated network design. A test using real world data from two French towns is presented in section 3 followed by a discussion and conclusions in section 4.

## 2. Network cartography from georeferenced points

Starting from a set of georeferenced points (manhole covers) and assuming the position of the network's general outfall, a methodology is put forward to retrieve the links between the points, i.e. the network pipes or edges. In a first, purely deterministic approach, the algorithm starts from the outfall point and chooses to link points with edges that minimize different cost functions. Secondly, this method is generalized with stochasticity introducing a probability based on the edges' costs allowing uncertainty propagation in the network generation.

We first assume that manhole covers are located accurately and are given as a set $S$ of georeferenced points $M_i(x_i, y_i)$, that constitute the nodes of the network to be mapped. Manhole localization may be done manually or automatically by using Very High Resolution aerial images as presented in [15, 34]. The general chart of the algorithm is presented in Figure 1 and the different steps

of the method are detailed in the following subsections. The full procedure is automated and written in Python language. Using the input points in ESRI shapefile format, it returns a new shapefile with an associated attribute table for the created segments.

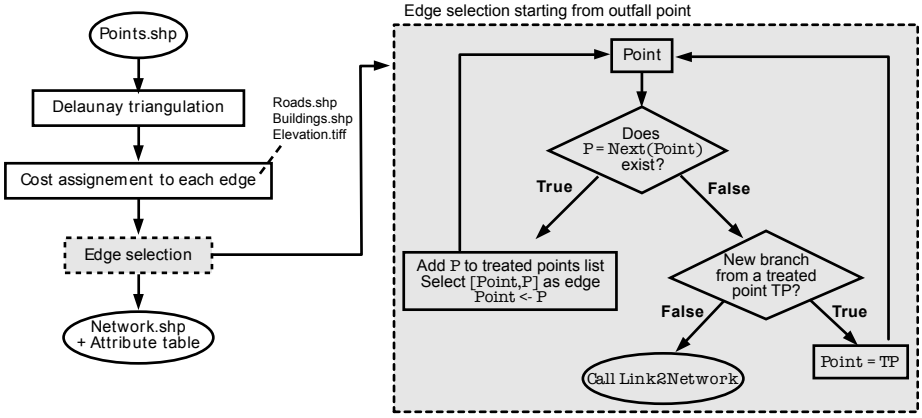



Figure 1: General chart (left) and edge selection algorithm (right). At the end of the process the `Link2Network` function is called for all remaining untreated points.

### *2.1. Creation of a valued directed graph*

In a first step, a Delaunay triangulation is computed from the set $S$. This operation consists in connecting the points of set $S$ to form triangles such as no point $M_i$ is inside the circumcircle of any triangle. In Euclidian distance, a Delaunay triangulation can be proven to be a spanning graph of set $S$. However, as this rule has no set rationale for wastewater networks, points located within a radius set by the user are also considered as possible neighbouring points. The edges thus formed give a consistent subset of all the possible connections between points of $S$, called initial graph in the following. Then, a value $c(M_iM_j)$ is assigned to each edge $M_iM_j$ which defines the "cost" to make wastewater flowing from $M_i$ to $M_j$. Of course, the edge is directed so that $c(M_iM_j) \neq c(M_jM_i)$.

At this stage, the cost function is based on two criteria: *i)* length: favouring

smallest edges, and *ii)* slope: favouring gravity fed flow within current industry rules used in France. If no other information than manhole cover position is considered available, the points elevation is assessed from a Digital Elevation Model (DEM) and assumed related to that of the pipes. This is a strong hypothesis, often used in hydrological modelling [9] but that is not entirely true for stormwater networks where ground slopes may be countered by adjusting the trench dimensions or adding pumping devices. Thus, water may flow in the opposite direction to surface flow/slope. However, in data scarce situations this approximation is retained as it is the only possibility. This assumption was verified by calculating the coefficient of determination between the DEM elevation and the pipe's inlet depths for the data we had access to. The results show that $R^2$=0.97 for Prades-Le-Lez. In the following, a third criterion based on the angle between two adjacent pipes will be added. When information on land use is available, which is the case in France via IGN BD-TOPO©, penalties are assigned to edges that cross roads ($P_r$) or buildings ($P_b$) using a buffer created around the roads' polylines. Note that similar criteria have also been used by [8] to generate a synthetic stormwater map.

The cost function is then defined as:

$$c(M_iM_j) = \alpha_L C_{L_{ij}} + \alpha_S C_{S_{ij}} + P_r + P_b \tag{1}$$

Where:

- $C_{L_{ij}}$ is the cost associated to the length $L_{ij}$ of edge $M_iM_j$. In Southern France, the maximal distance recommended between two manhole covers is about 80 m [5]. In the framework of network cartography with partial information on the manhole covers' position *e.g.* obtained through remote sensing data, the cost associated with the length increases linearly from 0 to 1 between 0 m and $L_{\max}$ (see Figure 2a), where $L_{\max}$ is chosen equal

to 160 m, in the following:

$$C_{L_{ij}} = \begin{cases} \frac{L_{ij}}{L_{\max}} & \text{if } L_{ij} < L_{\max} \\ 1 & \text{else} \end{cases} \tag{2}$$

Lmax is set to twice the recommended value as a tolerence to the older portions of networks which were designed before the fairly recent recommendation was made and to take into account non-reported manhole cover positions.

- $C_{S_{ij}}$ is the cost associated to the slope $S_{ij}$ of edge $M_iM_j$. The recommended slope to ensure gravity fed flow is between 0.3% and 0.7% and should not be greater than 5% nor negative (counterslope); the cost associated to the slope is thus defined as (see Figure 2b):

$$C_{S_{ij}} = \begin{cases} 0 & \text{if } 0.3\% < S_{ij} < 0.7\% \\ \frac{|S_{ij} - 0.7\%|}{10\% - 0.7\%} & \text{if } 0.7\% < S_{ij} < 10\% \\ -\frac{|S_{ij} - 0.3\%|}{1.3\%} & \text{if } -1\% < S_{ij} < 0.3\% \\ 1 & \text{if } -1\% > S_{ij} \; OR \; S_{ij} > 10\% \end{cases} \tag{3}$$

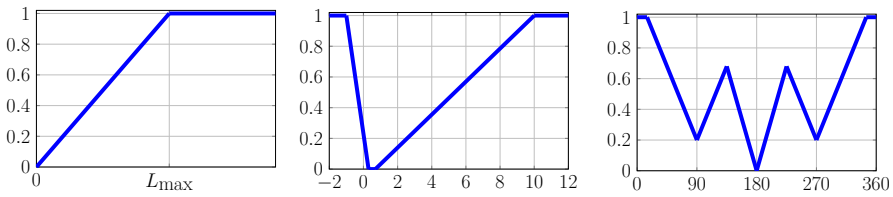
Note that the cost associated to the slope may be relaxed if the slope is estimated using ground elevation that may not be representative of underground elevation.

- $P_r$ is defined as the length of edge that is outside a road divided by a distance $d$. In this application $d$=20 m so that $P_r$ is greater than one if more than 20 m of an edge are outside the road buffer.

- $P_b$ is defined as the percentage of edge that crosses a building multiplied by a number $N$. In this application $N$=4, so that $P_b$ is greater than one

if more than 25% of the edge is crossing a building.

- The alpha coefficients are mere weighing coefficients. The user may set the alpha values without resorting to calibration simply by choosing the relative weight of each criterion. Alpha values reflect also the quality of the used datasets. For instance, if the slopes used to remap the network are known to be more accurate than the manhole cover locations which determine the pipe lengths, a higher alpha value may be given to slope alpha $\alpha_S$ and a lower value to the length alpha ($\alpha_L$) and vice versa. The alpha values may also be determined by referring to the local rules of practice and their evolution through time, or based on the development history of a given town. In France the rules of practice are based on national technical guidelines. At regional level, good practice rules may be established as well. Hence, there is little room for improvisation or modification for a given town. However, if for other countries these rules are set at regional or district level than the user may change them. If these guidelines change over time, it is also possible to modify the weighing of the parameters. If partial field data is available, they may also be evaluated by calibration or trial and error, provided the thresholds of the cost functions are modified accordingly.




(a) $C_L$ function of $L_{ij}$ (m) (b) $C_S$ function of $S_{ij}$ (%) (c) $C_\theta$ function of $\theta_{kij}$ (°)

Figure 2: Cost functions, related to edge length (left), slope (middle) and angle (right). Note that the angle cost function cannot be computed directly on the initial graph but is used during the pipe selection procedure, to chose between candidate edges.

This first step results in an directed graph with valued edges among which the network pipes will be selected. The attribute table of the generated shapefile contains useful information for each edge, such as point ID and elevation, edge length and slope, as well as cost values and road and building penalty values (see Table 1).

| Symbol | Definition |
|---|---|
| $M_u$ (resp. $M_d$) | Upstream (resp. downstream) point ID |
| $z_u$ (resp. $z_d$) | Estimate of upstream (resp. downstream) point elevation |
| $L_{ud}$ and $C_{L_{ud}}$ | Edge's length and associated cost |
| $S_{ud}$ and $C_{S_{ud}}$ | Edge's slope and associated cost |
| $P_{r_{ud}}$ | Road penalty |
| $P_{b_{ud}}$ | Building penalty |

Table 1: Attributes associated to the initial graph shapefile for each edge ($M_uM_d$).

The pipe selection algorithm (Figure 1 right) from the possible edges is applied to link the points of set $S$ starting from the outfall.

Cost function given by equation (1) is used to select the edge $ij$ starting from the current point $i$. At this step the following new criterion $C_{\theta_{kij}}$ is added. The coefficients 0.8 and 0.4 have been chosen to get the specific shape of the cost function plotted in Figure 2c, favouring 180° angles between two adjacent edges $ki$ and $ij$, and 90° angles to a lesser extent, and giving the highest penalty to acute angles:

$$C_{\theta_{kij}} = \begin{cases} 1 & \text{if } |360-\theta| < 30 \\ 0.8\frac{|90-\theta|}{60} + 0.2 & \text{if } |90-\theta| < |180-\theta| \\ 0.4\frac{|180-\theta|}{90} & \text{if } |180-\theta| < |90-\alpha| \text{ and } |180-\theta| \leq |270-\theta| \\ 0.8\frac{|270-\theta|}{60} + 0.2 & \text{if } |270-\theta| < |180-\theta| \end{cases} \tag{4}$$

The "angle cost" $C_{\theta_{ij}}$ of a given edge $ij$ is computed as the sum of the costs $C_{\theta kij}$ for all points $k$ already linked to point $i$. The total cost function $C(M_iM_j)$

given in equation (1) is thus modified as follows:

$$c(M_iM_j) = \alpha_L C_{L_{ij}} + \alpha_S C_{S_{ij}} + \alpha_\theta C_{\theta_{ij}} + P_r + P_b \tag{5}$$

Starting from the outfall, the edge with the lowest cost is chosen as pipe network provided that it is lower than a maximal admissible cost `MaxCost`. In the following `MaxCost` is set equal to one, so that, given the penalties on the road or building crossings, the points located on the roads are first linked together. All the linked points are labeled as `Treated Point` and stored in a list. When it is not possible to connect a new pipe to point $i$ with an admissible cost, the algorithm browses the list of already treated points to find a new path, called `Branch`, still favouring points located inside the road buffer. When all the possible branches have been explored, the algorithm browses the list of non treated points to connect them to the network using the `Link2Network` function with a relaxed constraint on road crossing. The pseudo code of the main treatment and of the Next function are presented in algorithms 1 and 2.

If some points are still unlinked at the end of this process, the point with the lowest elevation value is automatically defined as a new outfall, as stormwater networks might have more than one outlet on a given catchment. The linking operation starts over for all the remaining points (if any).

The entire process thus yields a multigraph with possible unlinked points corresponding to the false positives of a previous detection step. Although some authors consider these as "infeasable solutions" when optimizing network design [1], they do correspond to real-world situations where, due to urban growth, the network layout may be altered over time.

**Algorithm 1** Pseudo code of main treatment (gen stands for the NetworkGenerator class)

```
PtBranch ← [outfall]
while MainTREATMENT do
    BRANCH ← False
    idPt ← PtBranch[-1] # last point added to the list
    Next ← pointNext(gen, TreatedPt, idPt)
    if Next then # there is a Next on the same branch
        gen.NetWk.append(gen.edges[(Next, idPt)])
        gen.edges[(Next, idPt)].branchID = branchID
        TreatedPt.append(Next)
        PtBranch.append(Next)
    else
        # if no more point can be added, try to start a new branch from already treated points
        i = 0
        while (not BRANCH and i < len(TreatedPt)) do
            idPt ← TreatedPt[i]
            i ← i + 1
            Next ← pointNext(gen, TreatedPt, idPt)
            if Next then
                BRANCH ← True # New branch
                gen.NetWk.append(gen.edges[(Next, idPt)])
                TreatedPt.append(Next)
    if not BRANCH then # no new branch found
        # browse untreated points to link them to network
        couple ← Link2Network(gen, TreatedPt)
        if couple then
            TreatedPt.append(couple[0])
            gen.NetWk.append(gen.edges[(couple[0], couple[1])])
        else
            MainTREATMENT = False
```

### *2.2. Stochastic approach*

Each step of the previously developed process is subject to uncertainty: the cost function was defined from best practice recommendations given in the form of intervals that may not always be strictly followed; pipe slopes might not be reported accurately or could be determined using ground elevation data; maps of manhole positions might be erroneous or incomplete. Validation maps produced at the district or city scale may also have cartographic errors due to edge

**Algorithm 2** Pseudo code of function Next

```
NextPt ← NONE
Weight ← MaxCost
for Neigh in gen.neighbour[idPt] and not in TreatedPt do
    # Find DownPt, the downstream point of idPt and compute angle
    angleCost = gen.angleCost(DownPt, idPt, Neigh)
    # Find Neigh2, a possible upstream point of idPt
    for Neigh2 in gen.neighbour[idPt] do
        if idPt is the downstream point of Neigh2 then
            # add angle cost for all already present edges
            angleCost ← angleCost + gen.angleCost(Neigh2,idPt,Neigh)
    # Compute total cost
    gen.computeCost(gen.edges[(Neigh, idPt)])
    # keep the minimum cost (or use the probability in the stochastic version)
    if gen.edges[(Neigh, idPt)].cost < Weight then
        Weight ← gen.edges[(Neigh, idPt)].cost
        NextPt ← Neigh
return NextPt
```

matching and rubber-sheeting. It is thus admitted that the layout of the actual wastewater network cannot be retrieved perfectly. The aim of the process is then to propagate input data and rules' uncertainty and consequently map a set of probable network that explicitly represent the uncertainty in the wastewater networks accordingly. This enables to further propagate this mapping uncertainty into the hydraulic software to reproduce the main discharge uncertainty at outfall.

In this stochastic approach, for a given downstream node $M_i$ we define a probability distribution function $P_{M_iM_j}$ for each edge $M_iM_j$ of the initial graph, based on the edge's cost value:

$$P_{M_iM_j} = \frac{\frac{1}{c_{M_iM_j}}}{\sum_j \frac{1}{c_{M_iM_j}}} \tag{6}$$

Lower costs are assigned to the edges that meet the constraints imposed on length, slope and angle, hence their selection as pipe has a higher occurrence

probability. These correspond to a small-set of edges, which would have had a lower selection probability if the process was purely random.

For a given point $M_i$, instead of choosing the edge with the minimum cost, the upstream point $M_j$ is randomly sampled from possible neighbours in the initial graph using multinomial laws with parameters from equation (6). Several runs of the algorithm are performed so that, at the end of the process, a set of probable networks is thus obtained.

### *2.3. Validation procedure*

The mapped network is validated against the actual map via positional errors and changes in network hierarchy.

Positional errors are calculated using the criteria of completeness, correctness and quality, put forward in [18] to evaluate automatically extracted road networks, and total error is defined using [28]: [25]

$$\text{Completeness} = \frac{\text{TP}}{\text{TP+FN}} \tag{7a}$$

$$\text{Correctness} = \frac{\text{TP}}{\text{TP+FP}} \tag{7b}$$

$$\text{Quality} = \frac{\text{TP}}{\text{TP+FP+FN}} \tag{7c}$$

$$\text{Error} = \frac{\text{FN+FP}}{\text{RL}} \tag{7d}$$

where TP represents true positives (*i.e.* the length of correctly mapped pipes), FN false negatives (length of pipes that are not mapped by the algorithm but that do exist in the validation database), FP false positives length of pipes that are mapped by the algorithm but that are not reported in the validation

database) and where RL stands for Real Length *i.e.* the total length of pipes as reported in validation database. Completeness represents the percentage of the reference network which lies within the buffer around the mapped pipe. Correctness represents the percentage of correctly mapped pipes, i.e., the percentage of the mapped pipes which lie within the buffer around the reference network. Quality is a measure of the overall goodness of the final result. It takes into account completeness and correctness.

The optimum value is 1 for the three first criteria and is 0 for the error. In this application the buffer is flat and double sided. Its width is set to $w$ =5 m, *i.e.* ±2.5 m. This means that mapped pipes are considered to be identical to the real pipes if they fall within a 5 m strip. This distance may not meet the precision limits imposed by current legislation in developed countries (such as the U.S. Clean Water Act, or the French environmental legislation [22] revised in 2018). However, these maps may be used to plan database updating field work. In addition, in many developing countries, no maps exist at all and even a 5 m tolerance may be considered as an improvement.

Network hierarchy is based on Shreve's magnitude [38], which is a numerical measure of its branching complexity and implicitly accounts for topology. The network is assumed to be a mathematical tree and channels or branches, which have no upstream junctions or tributaries, have an order of 1. When two branches of order $i$ and $j$ join, the resulting downstream branch has an order $i + j$. Changes in network hierarchy will be directly reflected by the outlets order.

## 3. Application and results

The methodology is first tested on the town of Prades-le-Lez in South Eastern France (34°41'51"N;3°41'51"E) and validated on the town of Ramonville-

Saint-Agne in the South West of France (43°34'04"N;3°54'07"E).

The Prades-le-Lez database for the existing manhole locations is provided by Montpellier Méditerranée Métropole, the local operator. The town limits the manhole cover positions and the initial graph are presented in Figure 3. The database includes 792 manhole covers and 23.45 km of pipes with a mean length of 28 m and standard deviation of 15 m. All the other geographical data used in this study are available through the French National Institute of Geography's databases (BD-TOPO® and ALTI-RGE®). The network is not fully connected as can be seen on Figure 3 in the eastern part of the town.

The Ramonville-Saint-Agne database has been obtained through the French government's open access gateway (https://www.data.gouv.fr/fr/). Ramonville-Saint-Agne is a small town located in the urban area of Toulouse, the 4th biggest city in France. Despite its small size, it has a high population density, 2140,7 inhabitants.km$^{-2}$ in 2015 i.e. 13829 inhabitants over 6km$^2$ and is under constant urban pressure due to its proximity with the city of Toulouse. The network consists of 1878 nodes and 58.9 km of pipes, with a mean length of 30 m and standard deviation of 16 m.

The distance between the two towns is of 249 km. They are both part of the two most dynamic metropolitan areas of France, with a mean annual population growth of 2.9 % between 2010 and 2015 for Ramonville-Saint-Agne and 2.8% for Prades-le-Lez. In comparison, the city of Paris scores -0.3% over the same period and its metropolitan area 0.5%. Both catchments are larger and their networks are bigger than most of the study cases reported in the literature [3, 8, 31, 37].

### *3.1. Prades-le-Lez*

In both applications, the road buffer width for the road penalty $P_r$ is set to 8 m ($\pm$4 m). Using the 792 manhole cover locations, 9946 oriented edges

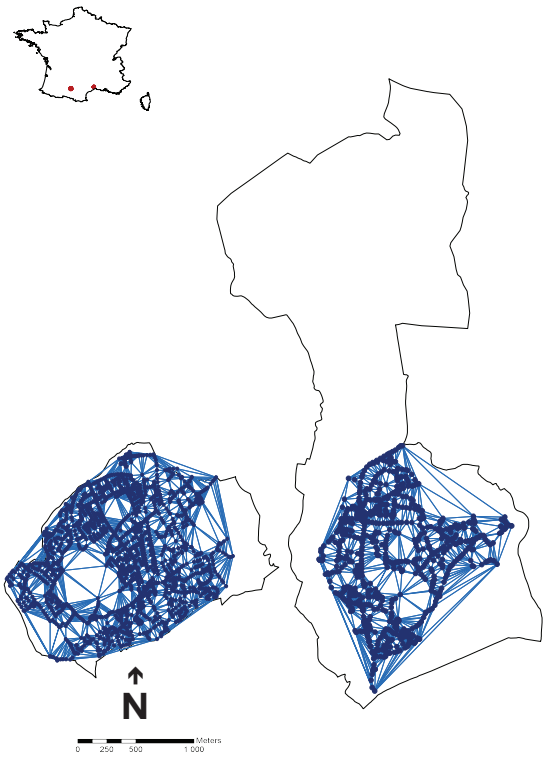



Figure 3: The two study catchments, Ramonville-Saint-Agne (Left) and Prades-Le-Lez (Right). Limits of the municipality, manhole cover positions as reported in the stakeholder's database and corresponding initial graphs.

are defined. The best ones, having the optimal cost values, will be selected as pies for the mapping process. In the following subsections, first the optimal $\alpha$ values are identified, then the "optimal" networks with regards to the various criteria are presented. Finally the set of probable networks obtained using the stochastic procedure is analyzed.

### *3.2. Selection of optimal $\alpha$ values*

A sensitivity analysis of the results is performed with respect to the three parameters $\alpha_L$, $\alpha_S$ and $\alpha_\theta$ which are varied from 0 to 1 by step of 0.1, while

respecting a total sum value of 1. Figure 4 represents the results in terms of error (eq. 7d).

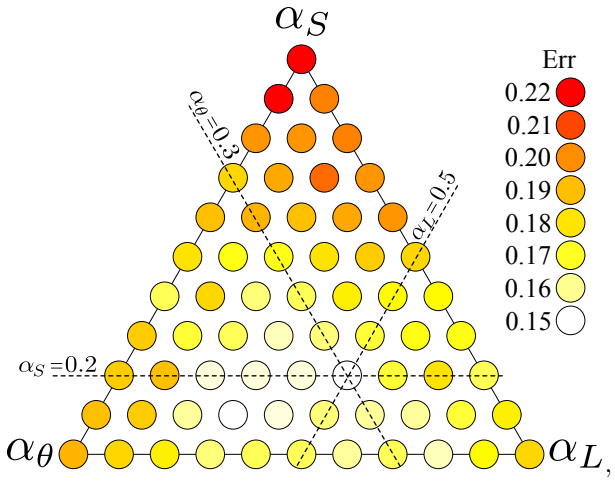
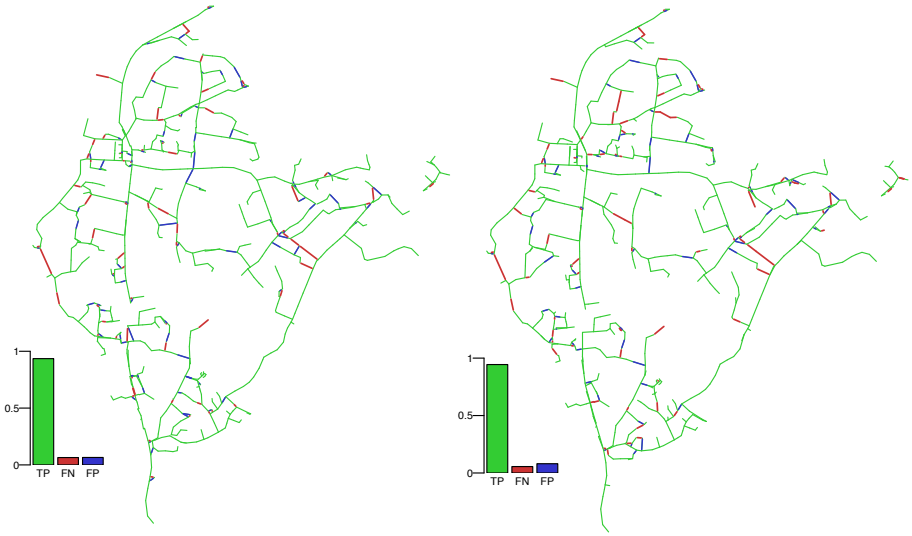



Figure 4: Results according to different values of $\alpha$ in term of error (eq. 7d). Each parameter $\alpha$ is equal to one at the corresponding top of the triangle and zero at the opposite base. The best results are obtained with $\alpha_L = 0.5$, $\alpha_S = 0.2$ and $\alpha_\theta = 0.3$.

The best result is obtained for $\alpha_L = 0.5$, $\alpha_S = 0.2$ and $\alpha_\theta = 0.3$, with an error equal to 0.158. This corresponds also to best quality (0.85) and correctness (0.92) and to a completeness value of 0.92. The resulting network is presented in Figure 5a.

Note that all four criteria have better values with this algorithm than when using the Kruskal algorithm with ranked error criteria on slope presented in [14]. Low $\alpha_S$ values would suggest that the slope is less important than distance when linking the network nodes which does not seem quite logical for a gravity fed flow network. A possible explanation is that we used terrain slope values, via the Digital Elevation Model, which are not representative of pipe slopes. Indeed, pipe layers often adjust the trench dimensions to imposed slope values through digging and filling. Thus, associating a cost function to the terrain slope values does not impact the simulation outputs greatly in terms of position.

Best completeness (0.93) is obtained with $\alpha_L = 0.2$, $\alpha_S = 0.2$ and $\alpha_\theta = 0.6$




(a) (b)

Figure 5: Best result: a) in term of error, quality and correctness, obtained with $\alpha_L = 0.5$, $\alpha_S = 0.2$ and $\alpha_\theta = 0.3$; b) in terms of completeness obtained with $\alpha_L = 0.2$, $\alpha_S = 0.2$ and $\alpha_\theta = 0.6$.

(Figure 5b). The higher $\alpha_\theta$ and lower $\alpha_L$ values result in a network that has more right angle connections but edges of similar length: in comparison with the best error results, the difference between the minimal edge values are of 0.1 m, while the maximum values are 0.6 m different. Given the positional errors and the precision of the data, these differences are thought to be not significant. There are however 6 non-connected manhole covers when angles are given a higher weight compared to 2 when distance is given a higher weight.

In both instances isolated pipe segments can be found, highlighting the fact that the proposed procedure allows for disconnected graphs in order to reproduce sub-networks, such as the one located in the eastern part of the catchment. However, with lower $\alpha_L$ values the number of disconnected parts increases from 2 to 5. These are not false connections but rather incomplete connections as can be seen when overlaying the simulated networks maps to the actual map.

Additional dissimilarities with the ground truth (see Figure 6 and following subsection) also rise from the fact that the mapping algorithm considers that flow is gravity fed while the real network has a mainforce in the western part of the catchment.

### *3.3. Network connectivity and layout*

In terms of network layout and connectivity, the maximum Shreve's magnitude of both the "Best Network" and the "Most Complete" network are within the same order of magnitude as the real network's: Shreve's magnitude at the outlet is of 131 for the actual network, 134 for the best network and 149 for the most complete network (see Figure 6). However, while the real network has its highest orders on the western part of the catchment, both simulated networks have higher orders in their eastern branch. This is because the error criteria we have used to assess the results, rely only on position and not on flow direction. Indeed connections are allowed even for pipes with counter slopes. Thus 43% of the connected pipes have a slope cost value of 1 for the "Best Network" and 39.7% for the "Most complete" network. All of these pipes are true positives and do exist in the validation database and 59% of them have reported negative slope values.

To impose correct flow directions in the network, solutions with higher alpha $\alpha_s$ values should be selected. Our results however indicate that all solutions with $\alpha_s \geq 0.7$ and $\alpha_L < 0.2$ fail to connect the upstream and downstream parts of the network resulting in lower Shreve's magnitudes at the outlet and networks with less spatial extent. The corresponding Shreve's magnitudes vary between 33 and 59. Given the catchment's topographic configuration, relaxing the condition on gravity fed flow is the only possibility of insuring connectivity throughout the network. However, none of our solutions is able to reproduce the impact of the

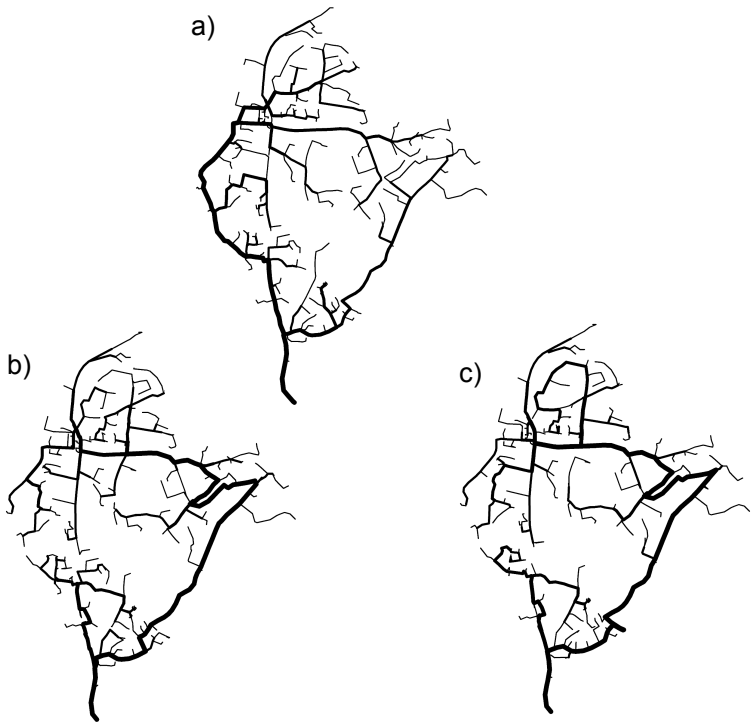



Figure 6: Comparison of Shreve's number between the actual and mapped networks. a) actual network; b) best mapped network obtained in terms of error, quality and correctness; c) best mapped network obtained in terms of completeness.

mainforce and to insure flow in the lower western part of the catchment. Such a-priori knowledge may be translated by assigning an overall null cost value to the corresponding edges thus forcing the connection algorithm to select them. The same procedure may be used if information is also available on the location of pumping stations.

### *3.4. Uncertainty propagation*

Using best $\alpha$ parameters determined previously, the algorithm in its stochastic version is run 500 times. When some edges are selected either in one direction or in the other, the two occurrence probabilities are summed, as no information is available in the validation database on the actual network topology and flow

directions. Thus 1570 non oriented pipes are selected from the 9946 edges of the initial graph. The results are plotted in Figure 7.

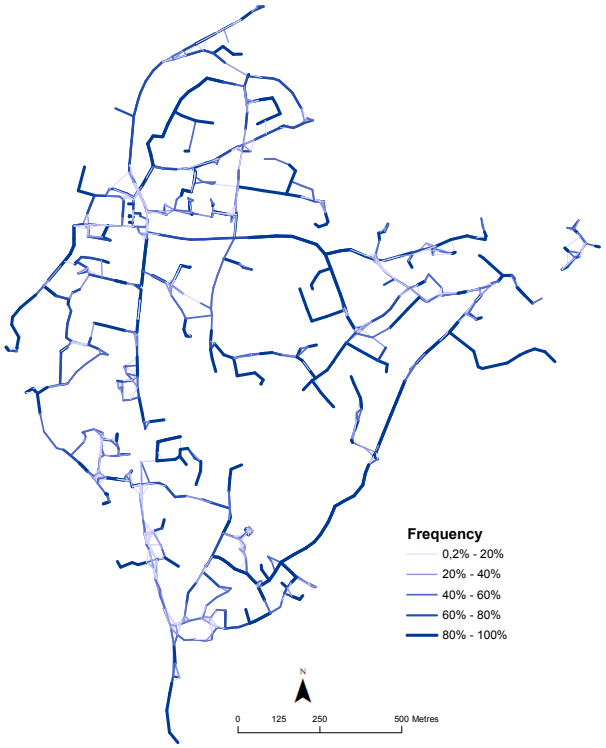



Figure 7: Results of uncertainty propagation. Darker lines correspond to pipes with higher occurrence frequency.

For the sake of clarity, only pipes selected at least for one network are represented, according to their occurrence frequency in the 500 runs. 708 of these pipes are reported in the validation database and 862 are not. The occurrence probabilities of these True and False pipes are plotted in Figure 8 in boxplot format.

The results indicate that the True Positive distribution median is equal to 86%, meaning that 50% of pipes of the validation database are selected by the stochastic process in more than 86% of the 500 simulations. More than 25% are

selected each time and 25% have an occurrence frequency lower than 55.95%. Half of the false pipes have an occurrence frequency lower than 16.8%.

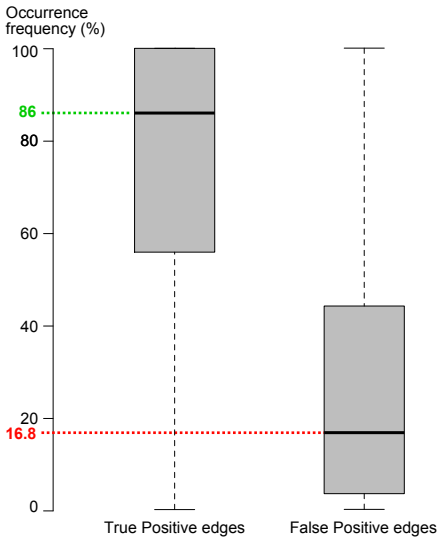



Figure 8: Results of the uncertainty propagation. Frequencies of True Positive and False Positive pipes

It can be seen in Figure 7 that the pipes with the highest occurrence probability (90% to 100%) correspond to the starting segment of the outlet and to the branches with no tributaries that are the outer pipes of the network. The median length of these pipes is 31.2 m (mean = 32.8 m; $\sigma$=15.7; N=332). The pipes with the lowest occurrence probability (0.2% to 20%) are those linking nodes with very similar elevation values or that have counter slope sections. On average they are longer with a median length of 41.2 m (mean length=42.7 m, $\sigma$=18.7, N=737).

38 of the 8376 non selected edges are reported in the validation database and 34 of them have a prohibitive $C_L$ value as they fall outside the road buffer, some run in fact parallel to the buffer (see Figure 9).

The results of the uncertainty propagation show that, in our case study, half of the generated network pipes are certain. They also show that about 85 % of

Figure 9: Examples of initial graph edges that are never selected as pipes while being part of the validation database.

graph edges linking manhole cover locations are certain not to host a network pipe. These results are however case study and input data dependent.

## 4. Validation

In order to test the robustness of the method we randomly sampled 75%, 50% and 25% of the manhole positions i.e. 592, 395 and 198 points respectively. We ran the algorithm again with the parameters corresponding to the lowest error, best quality and correctness i.e $\alpha_L = 0.5$, $\alpha_S = 0.2$ and $\alpha_\theta = 0.3$. When omitting intermediate manhole covers, the possibility of crossing roads and buildings increases, as manholes are mandatory at pipe junctions and cross-

roads. The cost function value associated to length, slope and angles may also increase. The `MaxCost` parameter is thus taken respectively equal to 1.5, 2 and 4 for the 75%, 50% and 25% samples. The results are presented in figure 10 and show that:

- When using 75% of the positions, 581 pipes are created and 2 points remain unconnected. The positional errors increase with completeness dropping to 0.79 while quality and correctness reach 0.6 and 0.72 respectively. The overall error is of 0.51.

- When using 50% of the positions, 379 pipes are created and 3 points are left unconnected. The error increases to 1.05 and correctness, quality and completeness drop to 0.48, 0.37 and 0.63 respectively.

- When using 25% of the manhole cover locations 196 pipes get created and no points are left unconnected. The resulting network has very low quality criteria (error = 1.87, Correctness = 0.23, Quality = 0.17, Completeness = 0.37) but is not fragmented because of the increase of the `MaxCost` parameter.

These results show that the outcome of our method is dependent on the density of points used to generate the network. If less than half of the manhole positions are detected through field data or photo-interpretation, the algorithm will fail to reproduce a map of acceptable quality. However, as can be seen in Figure 10, the overall layout of the network is still visible and the main trunks are mapped even when using 25% of the original dataset.

A second validation test consisted in running the algorithm on the town of Ramonville-Saint-Agne. We ran the algorithm again with the same parameters i.e $\alpha_L = 0.5$, $\alpha_S = 0.2$ and $\alpha_\theta = 0.3$.

The corresponding error is of 0.28, quality is 0.76, correctness is 0.84 and com-

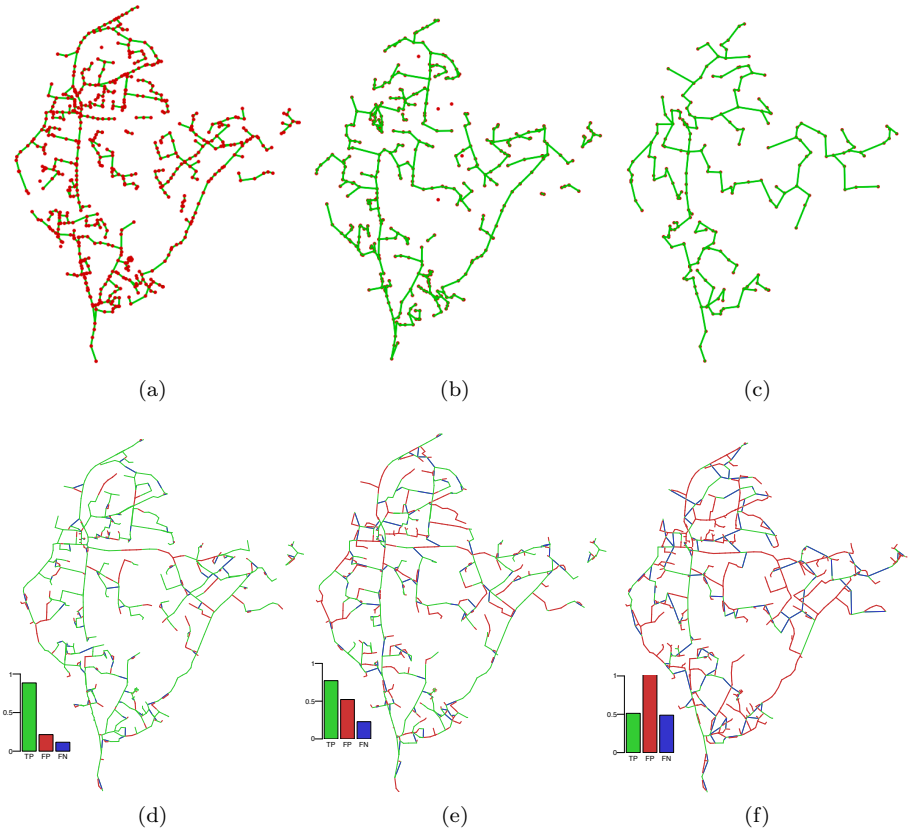



Figure 10: Results obtained with 75% (a,d), 50% (b,e) and 25% (c,f) of the database points. Up: initial points (red) and mapped network (green), down: validation against real network.

pleteness 0.89. 1817 pipes are mapped by the algorithm and there are only 32 un-connected manhole covers 30 of which are located outside the road buffer and are less likely to be selected by the algorithm. This is a common problem as 50.2 % of the false negatives, i.e. pipes that do exist in the validation database but that the algorithm did not map, are located outside the road buffer. 24% correspond to short pipes ($< 20$ m) and 1% to pipes $>80$ m. The proportions are similar for the false positives, with 40% of short pipes and 0.9% of pipes with a length $>80$ m. 30% of the false positive have acute angles and 42% have a slope cost The results of the sensitivity test on Ramonville-Saint-Agne pre-

sented in Figure 11. They show that three triplets give the best results in terms of lowest error and highest completeness (0.89), correctness (0.85) and quality (0.77) : $\alpha_L = 0.8$, $\alpha_S = 0.0$ and $\alpha_\theta = 0.2$; $\alpha_L = 0.8$, $\alpha_S = 0.1$ and $\alpha_\theta = 0.1$ and $\alpha_L = 0.7$, $\alpha_S = 0.1$ and $\alpha_\theta = 0.2$. The range of error values calculated on this catchment is higher than those calculated on Prades-Le-Lez. However using lower coefficients for slope i.e. lower $\alpha_S$ values, still yields better results.

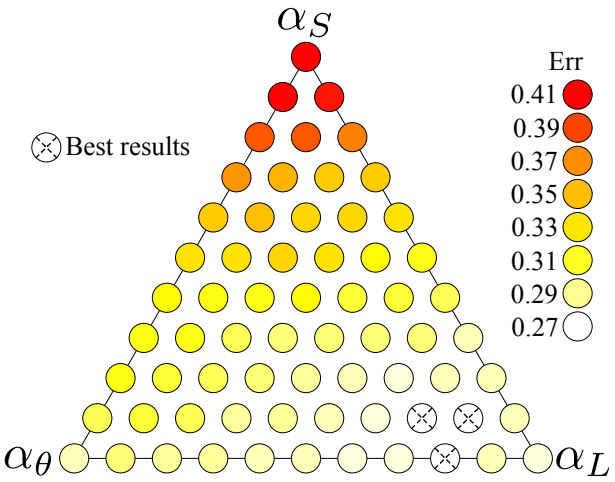
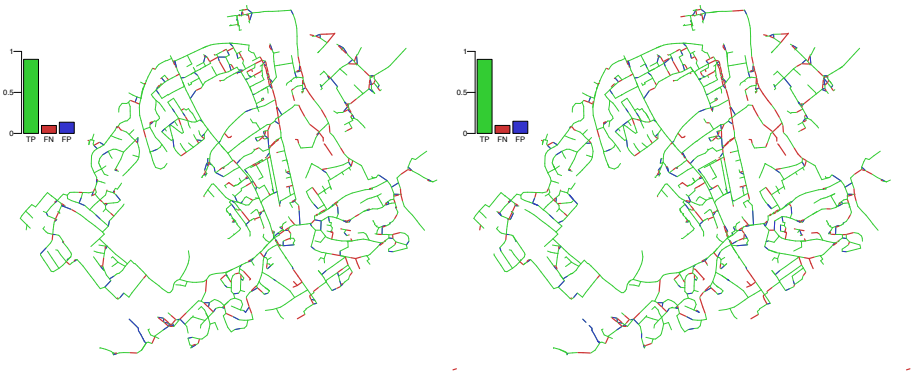



Figure 11: Error values according to different values of $\alpha$ for Ramonville-Saint-Agne (eq. 7d).

Figure 12 shows the networks generated using both the parameters found for Prades-Le-Lez and with the first triplet. The higher influence of distance vs slope is again thought to be due to the use of surface elevation data which is not necessarily representative of the pipe slopes : 43 % of the true positives in figure 12b have a slope cost function value of 1. Unfortunately 92% of the slope values are missing in the validation database. The remaining 8% have a mean slope value of 2% ($\sigma$= 2.56%; N=95).

In terms of connectivity however, the results are very poor. This is due to the fact that, on the one hand, Ramonville-Saint-Agne's actual network is not fully connected according to the digital maps we obtained despite the existence of three mainforces. On the other hand, as highlighted previously, the DEM




(a) (b)

Figure 12: Results obtained on Ramonville-Saint-Agne with a) $\alpha_L = 0.5$, $\alpha_S = 0.2$ and $\alpha_\theta = 0.3$ ; b) $\alpha_L = 0.8$, $\alpha_S = 0.0$ and $\alpha_\theta = 0.2$.

data indicates that more than 40 % of the true positives, i.e. pipes that do exist in reality have counter slopes. Hence, our algorithm creates 308 network branches but the highest Shreve order we calculate is 7 while the sub-network of the validation database has a Shreve order of 387.

## 5. Discussion and conclusion

In this work we put forward a method to build a map of an underground network using a set of georeferenced points. The method has been developed to help contractors, managers or scientists infer the location of the wastewater network and may be used at the scale of any administrative entity: borough, town or water board where information is missing.

The method we developed requires less input data than some of the methods described in the literature which consider the network layout as an optimization problem based on pipe size or network cost [2, 20, 40, 31, 37]. However, the connectivity data we have produced in this work is not enough to run a hydraulic or hydrological model of the network. A procedure to recreate the networks full geometry *i.e.* pipe diameter, rim and invert elevation and depth,

and insure proper conveyance is currently under development with a method that does not involve any calibration to determine these values and relies only on hydraulic continuity rules and industry practices. Similarly to the approach followed by [3] we use scripts to propagate pipe attribute data. The code is currently being validated. The main challenge is to translate these industry practices into cost functions as we have done in this work. Note that the functions are generic and the thresholds may be changed according to the practices of a given country. Thus our algorithm may be used to generate the topology of the network which may be used in hydraulic modelling or integrated in asset management databases. Though the later should ideally contain information on pipe location, geometry and condition, in many developing and developed countries they do not exist or have very fragmentary data. Pipe condition and age are not always reported for instance. It would be presumptuous on our part to claim that we would be able to generate such information using only manhole cover positions. We are currently addressing this issue using text-mining and NLP techniques [12].

The algorithm presented in this work is very efficient in terms of computation time as only 6 min are required to build a single network of 792 nodes and calculate the corresponding validation criteria, using a standard PC (INTEL CORE I7-5930K, 3.5GHz with 16Gb RAM). For the stochastic approach, the Delaunay graph being built once for all, the time is reduced to about 4 min per probable network.

Two types of information are necessary to run our algorithm: the location of the network nodes *i.e.* manhole covers and the slope of the underlying pipes. This is the geometric feature which is most difficult to determine at this stage. It is also the most sensitive parameter both to control gravity fed flow and to model it. In some examples [2, 21, 27, 40], slope has been calibrated according to

discharge and velocity. This procedure is not indicated in our case because the main point of mapping the wastewater network is to use it in urban hydrological modelling to simulate discharge values based on measured precipitation data and not pre-determine them. Slope is often one of the least documented variables in French urban databases available in open access. In the case of Prades-le-Lez for instance, only 46% of the records have slope values and most of them are incoherent. This complicates both the validation process and the attempts made to find a rule or a method to predetermine slope values. Assuming that terrain and underground slopes are parallel would be a suitable hypothesis as long as there are no local constraints such as bedding planes or obstacles like other buried networks, which could result in different layouts. Despite our best efforts, we could not find any open access geotechnical reports, which could give us an indication on bedding planes. One solution would be to infer slope values based on the ground elevation and laying depths, knowing that in certain instances the laying depth is modified to insure adequate slope values and flow conditions. A linear relationship was established between the roadway elevation and the pipe's upstream and downstream inlet depths ($R^2$=0.97) for Prades-le-Lez. This result is promising and the method will be tested on other catchments. It does require accurate fine scale elevation data but a growing number of towns and county councils are providing this type of information through their open data web sites. In France for instance, based on the modified law n2016-1321 of October 7 2017, all local authorities employing more than 50 agents have to put administrative documents and public data online by October 2018. The lack of accurate data on pipe slopes may be compensated by information on the location of pumps or mainforces. The network's topology may be corrected by assigning an overall null cost value to the corresponding edges thus forcing the connection algorithm to select them. The same procedure may be used

if the information on flow directions is available. This algorithm is mainly developed for combined sewer outflows. It may also be used to map the layout of stormwater networks provided inlet grates are used as additional location indicators. The validation dataset we have used is one of the largest used in the literature to validate network mapping algorithms. The low error values and high scores for completeness, correctness and quality indicate that the method is robust and may be adapted and tested on other study zones.

## Acknowledgments

This study is part of the project Cart'Eaux funded by the European Regional Development Fund (ERDF). The authors would like to acknowledge the IGN (French National Geographical Institute), Montpellier Métropole and SICOvAL (Syndicat Intercommunal pour l'aménagement et le développement des coteaux et de la vallée de l'Hers) for access to their digital databases. They also would like to thank the three anonymous reviewers whose comments helped greatly improve the quality of the manuscript.

## References

[1] M. Afshar. Improving the efficiency of ant algorithms using adaptive refinement: Application to storm water network design. *Advances in Water Resources*, 29(9):1371–1382, sep 2006.

[2] M. Afshar. Partially constrained ant colony optimization algorithm for the solution of constrained optimization problems: Application to storm water network design. *Advances in Water Resources*, 30(4):954–965, 2007.

[3] M. Aguilar and R. Dymond. Innovative Technologies for Storm-Water Management Programs in Small Urbanized Areas. *Journal of Water Resources Planning and Management*, 140(11):04014029, 2013.

[4] A. Allard, K. Chancibault, and H. Andrieu. Construction of drainage network of an entire urban area - Case of Nantes East. In *Novatech 2013*, pages 1–10, Lyon, 2013.

[5] Aqua Valley. Charte qualité des réseau d'eau potable et d'assainissement en Languedoc-Roussillon. Technical report, AquaValley, 2017.

[6] J.S. Bailly, F. Levavasseur, and P. Lagacherie. A spatial stochastic algorithm to reconstruct artificial drainage networks from incomplete network delineations. *International Journal of Applied Earth Observation and Geoinformation*, 13(6):853–862, 2011.

[7] F. Bazlamaçci and K. Hindi. Minimum-weight spanning tree algorithms A survey and emphirical study. *Computers and Operations Research*, 28(8):767–785, 2001.

[8] R. Bertsch, V. Glenis, and C. Kilsby. Urban flood simulation using synthetic storm drain networks. *Water (Switzerland)*, 9(12), 2017.

[9] K. Beven. *Rainfall-Runoff Modelling: The primer*. John Wiley & Sons, Chichester, 2001.

[10] M. Bilal, W. Khan, J. Muggleton, E. Rustighi, H. Jenks, S. Pennock, P. Atkins, and A. Cohn. Inferring the most probable maps of underground utilities using Bayesian mapping model. *Journal of Applied Geophysics*, 150:52–66, 2018.

[11] F. Blumensaat, M. Wolfram, and P. Krebs. Sewer model development under minimum data requirements. *Environmental Earth Sciences*, 65(5):1427–1437, 2012.

[12] T. Bonnabaud La Bruyère, N. Chahinian, C. Delenne, L. Deruelle, M. Derras, F. Frontini, R. Panckhurst, M. Roche, L. Sautot, and M. Teis-

seire. Mégadonnées, données liées et fouille de données pour les réseaux d'assainissement. In M. Brando, C., Frontini, F., Roche, editor, *Atelier Humaintés Numériques Spataialisées (HumaNS'2018)*, pages 15–19, 2018.

[13] M. Brazil, R. Graham, D. Thomas, and M. Zachariasen. On the history of the Euclidean Steiner tree problem. *Archive for History of Exact Sciences*, 68(3):327–354, 2014.

[14] B. Commandré, N. Chahinian, J.S. Bailly, M. Chaumont, G. Subsol, F. Rodriguez, M. Derras, L. Deruelle, and C. Delenne. Automatic reconstruction of urban wastewater and stormwater networks based on uncertain manhole cover locations. In *14th IWA/IAHR International Conference on Urban Drainage, ICUD 2017*, volume 2012, 2017.

[15] B. Commandre, D. En-Nejjary, L. Pibre, M. Chaumont, C. Delenne, and N. Chahinian. Manhole cover localization in aerial images with a deep learning approach. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, 42(1W1):333–338, 2017.

[16] B. Golden, M. Ball, and L. Bodin. Current and future research directions in network optimization. *Computers and Operations Research*, 8(2):71–81, 1981.

[17] R. L. Graham and P. Hell. On the History of the Minimum Spanning Tree Problem. *Annals of the History of Computing*, 7(1):43–57, 1985.

[18] O. Heipke, C., Mayer, H., Wiedemann, C., Jamet. Evaluation of automatic raod extaction. *International Archives Photogrammetry and Remote Sensing*, 32:47–57, 1997.

[19] F. K. Hwang and D. Richards. Steiner tree problems. *Networks*, 22(1):55–89, 1992.

[20] S. Ivić, L. Grbčić, and S. Družeta. Cooperative Random Walk for Pipe Network Layout Optimization. *International Journal of Applied Engineering Research*, 11:2839–2847, 2016.

[21] J. Izquierdo, I. Montalvo, R. Pérez, and V. S. Fuertes. Design optimization of wastewater collection networks by pso. *Computers and Mathematics with Applications*, 56(3):777–784, 2008.

[22] JORF. Arrêté du 15 février 2012 pris en application du chapitre IV du titre V du livre V du code de l'environnement relatif à l'exécution de travaux à proximité de certains ouvrages souterrains, aériens ou subaquatiques de transport ou de distribution. *Journal Officiel de la République Française*, 45:2988, 2012.

[23] J. Kruskal. On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem. *Proceedings of the American Mathematical Society*, 7(1):48–50, 1956.

[24] N. Metje, P. Atkins, M. Brennan, D. Champan, H. Lim, J. Machell, J. Muggleton, S. Pennock, J. Ratcliffe, M. Redfern, C. Rogers, A. Saul, Q. Shan, S. Swingler, and A.N. Thomas. Mapping the underworld: State of the art review. *Tunnelling and underground space technology*, 22:568–586, 2007.

[25] du Tourisme et de la Mer Ministère de l'Equipement, des Transports, du logement. Fascicule n 70 Ouvrages dassainissement Titre I : Réseaux. *Bulletin officiel*, page p 149, 2003.

[26] M. Möderl, D. Butler, and W. Rauch. A stochastic approach for automatic generation of urban drainage systems. *Water Science and Technology:*

*a journal of the International Association on Water Pollution Research*, 59(6):1137–43, 02 2009.

[27] R. Moeini and M. H. Afshar. Layout and size optimization of sanitary sewer network using intelligent ants. *Advances in Engineering Software*, 51:49–62, 2012.

[28] I. Molloy and T. F. Stepinski. Automatic mapping of valley networks on Mars. *Computers and Geosciences*, 33(6):728–738, 2007.

[29] M Moy de Vitry, K Schindler, J Rieckermann, and Leit ao JP. Sewer inlet localization in uav image clouds: Improving performance with multiview detection. *Remote Sensing*, 10(7):706, 2018.

[30] J. M. Muggleton, M. J. Brennan, and Y. Gao. Determining the location of buried plastic water pipes from measurements of ground surface vibration. *Journal of Applied Geophysics*, 75(1):54–61, 2011.

[31] K. Navin and Y. Mathur. Layout and Component Size Optimization of Sewer Network Using Spanning Tree and Modified PSO Algorithm. *Water Resources Management*, 2016.

[32] J. Nešetil, E. Milková, and H. Nešetilová. Otakar Borvka on minimum spanning tree problem: Translation of both the 1926 papers, comments, history. *Discrete Mathematics*, 233(1-3):3–36, 2001.

[33] H. Niigaki, J. Shimamura, and M. Morimoto. Circular object detection based on separability and uniformity of feature distributions using bhattacharyya coefficient. In *21st Int. Conf. on Pattern Recognition (ICPR)*, 2012.

[34] J. Pasquet, T. Desert, O. Bartoli, M. Chaumont, C. Delenne, G. Subsol, M. Derras, and N. Chahinian. Detection of manhole covers in high-

resolution aerial images of urban areas by combining two methods. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(5):1802–1807, 2016.

[35] Population Division of the Department of Economic and Social Affairs of the United. World Urbanization prospects: The 2018 Revision, 2018.

[36] R. C. Prim. Shortest Connection Networks And Some Generalizations. *Bell System Technical Journal*, 36(6):1389–1401, 1957.

[37] G. Rezaei, M. Afshar, and M. Rohani. Layout optimization of looped networks by constrained ant colony optimisation algorithm. *Advances in Engineering Software*, 70:123–133, 2014.

[38] R. L. Shreve. Statistical law of stream numbers. *The Journal of Geology*, 74(1):17–37, 1966.

[39] R. Sitzenfrei, M. Möderl, and W. Rauch. Graph-based approach for generating virtual water distribution systems in the software vibe. *Water Science and Technology: Water Supply*, 10(6):923–932, 2010.

[40] G. A. Walters. The design of the optimal layout for a sewer network. *Engineering Optimization*, 9(1):37–50, 1985.